Model fitting using the least volume criterion

Chris Tofallis

University of Hertfordshire Business School
Dept. of Statistics, Economics, Accounting and Management Systems
Mangrove Rd, Hertford, SG13 8QF, UK
c.tofallis@herts.ac.uk

Abstract

Given data on multiple variables we present a method for fitting a function to the data which, unlike conventional regression, treats all the variables on the same basis i.e. there is no distinction between dependent and independent variables. Moreover, all variables are permitted to have error and we do not assume any information is available regarding the errors. The aim is to generate law-like relationships between variables where the data represent quantities arising in the natural and social sciences. Such relationships are referred to as structural or functional models. The method requires that a (monotonic) relationship exists; thus, in the two variable case we do not allow cases where there is zero correlation. Our fitting criterion is simply the sum of the products of the deviations in each dimension and so corresponds to a volume, or more generally a hyper-volume. One important advantage of this criterion is that the fitted models will always be units (i.e. scale) invariant. We formulate the estimation problem as a fractional programming problem. We demonstrate the method with a numerical example in which we try and uncover the coefficients from a known data-generating model. The data used suffers from multicollinearity and there is preliminary evidence that the least volume method is much more stable against this problem than least squares.

1 On the undeserved ubiquity of least squares regression

In fitting a function to data, conventional regression requires one variable to be 'special' — this is the dependent variable. In the sciences however, one often wishes to re-arrange the model equation by changing the subject of the formula. Statisticians tell us that in that case we should carry out a second regression. Yet scientists are uncomfortable with having separate models for each variable, which are not equivalent to each other and yet are meant to represent the same relationship. Calibration is another case where one would like mutual equivalence: e.g. in psychology one can have two tests that are intended to measure the same ability: a formula or conversion table is required to relate the score on one test to that on the other.

Another case where regression is inappropriate is where one wants to deduce a parameter such as the rate of change (slope). If both variables are subject to error then ordinary least squares will under-estimate the slope, and regressing x on y will over-estimate it. A simple example involves plotting galaxy speed (or redshift) against distance from the observer. The slope of the fitted line gives what is called the Hubble constant, whose

value crucially determines the future of the universe: will it continue expanding or will it eventually begin to collapse in on itself? Conventional regression gives different values for the Hubble constant depending on which variable is treated as being dependent, but there is no apparent reason for choosing one variable as against the other.

An oft-cited reason for using least squares fitting is that under certain assumptions on the errors, it will provide the best linear unbiased estimate ('BLUE') of the slope. This is the Gauss-Markov theorem, where 'best' is taken to mean minimum variance. What is not widely appreciated is that 'linear' here refers not to the form of the fitted model, but rather that the expression for the estimated coefficient be linear in y. One can find estimators with even lower variance by removing this non-essential condition e.g. other L_p-norm estimators are not linear in y.

In multiple regression it is widely, and mistakenly, believed that that the fitted coefficients tell us the contribution that a particular variable makes to the dependent variable. In fact, not even the sign of the coefficient can be relied upon to tell us the direction of the relationship i.e. a particular x-variable may be positively correlated with the y-variable, and yet have a negative coefficient in the regression model. This is the problem of multicollinearity: if there are near-linear relations among the explanatory variables then the coefficients produced by regression will not only be highly uncertain (large standard error) but also not be open to sensible interpretation.

We shall present a technique for model-fitting which treats all variables on the same basis. The method has the important property of being units-invariant; this is an advantage not shared by the total least squares approach (also known as orthogonal regression), and arises from the fact that we use the product of the deviations in each direction rather the sum (or sum of squares) when calculating the fitting criterion.

2 The least areas criterion

Consider a set of data points in two dimensions as in Figure 1. By drawing the vertical and horizontal deviations from the line we create a right-angled triangle for each data point. Our fitting criterion is simply to minimise the sum of these areas. A key advantage of this approach is that changing the units of measurement will not affect the resulting line. In other words it is a scale invariant method. Furthermore we can add a constant to either variable and the geometry is such that the line merely gets shifted vertically or horizontally. Combining the scale and translation invariance implies that the least areas line is invariant to linear transformations of the data. It is also apparent that switching the axes has no effect: the variables are treated symmetrically. (A textbook discussion of this method appears in Draper and Smith [5].)

We note that it is essential that there be a non-zero correlation in the data otherwise the method fails. For those seeking to quantify relationships between data variables in the experimental sciences this would hardly seem to be a restrictive requirement. However for those working in the area of design and who are concerned with geometrical shapes, it does rule out the fitting to data scattered around a vertical or horizontal line, or circle, or rectangle with sides parallel to the co-ordinate axes etc.. We shall not discuss fitting curves in this paper but we note that this method is not suitable for fitting a relationship

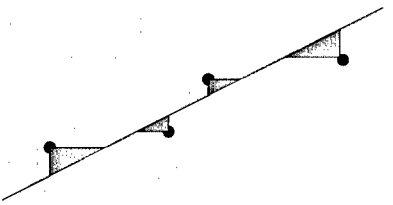

FIG. 1. Sum of areas to be minimised in least area calculation.

that is not monotone over the range of the data i.e. there cannot be maxima or minima over the data range otherwise the area deviation associated with a given point may not be uniquely specified. Such problems may be avoided by breaking up the data set into subsets at the optima and fitting a monotone function to each subset, thus producing a piecewise monotone function.

The least areas method has an interesting history, it has surfaced under different guises in diverse research literatures throughout the twentieth century. In astronomy it is known as Stromberg's impartial line. In biology it is the line of organic correlation. In economics it is the method of minimised areas or diagonal regression. In statistics it is sometimes referred to as the 'standard or reduced major axis'. This derives from the fact that if the data are standardised by dividing by their standard deviation, then the fitted line corresponds to the major (i.e. principal) axis of the ellipse of constant probability for the bivariate normal distribution. Yet another name for this technique is the geometric mean functional relationship. This is because the slope has a magnitude equal to the geometric mean of the two slopes arising from ordinary least squares (OLS) (proved in Barker, Soh and Evans [2], and Teissier [20]) i.e. if we regress y on x and get a slope b_1 and then regress x on y (so as to minimise the sum of squared deviations in the x- direction) and obtain a regression line $y = a + b_2 x$, then the geometric mean slope is $b = (b_1 b_2)^{1/2}$. It is interesting to note that the two OLS slopes are connected via the correlation between the variables

$$r^2 = \frac{b_1}{b_2}.$$

This implies that as the correlation falls the disagreement between the two OLS slopes increases; for example, even with a correlation as high as 0.71 one of these slopes will be twice as large as the other! It also follows that the magnitude of the slope of the least areas line lies between those of the two OLS lines. This is intuitively satisfying in a technique that aims to treat x and y deviations symmetrically. Specifically, for the case of positive but imperfect correlation, we have $b_2 > b > b_1$ because $b/r > b > rb$.

From the geometric mean property and the expressions for OLS slopes one can deduce that the magnitude of the slope of the least areas line takes on a particularly simple closed form: it is the standard deviation of y divided by the standard deviation of x. The sign of the slope is provided by the sign of the correlation between y and x.

Numerical experiments have been carried out to compare this fitting technique against five others (Babu and Feigelson [1]). A specified underlying model was used to generate data (mostly bivariate normal samples) and the aim was to see which method could

best recover the slope of the model. The simulations involved varying the sample size, correlation and variances. Orthogonal regression gave the poorest accuracy. There were two methods that came out with highest accuracy: the least areas method and the least squares bisector. The latter bisects the smaller angle formed between the two OLS lines. Unfortunately the OLS bisector is not units invariant and so does not suit our purposes (Ricker [17]).

Turning now to applications, the method seems to have appeared most often in the field of biometrics (the application of statistics to biological data). For example, in relating the size of one body part to another (or to the total weight or height) in humans and other animals, one may collect data from an individual at successive stages in their growth, or from many individuals at different points in their development. It is not generally possible to distinguish between dependent and independent variables in such a context. Isometric growth is the special case where the two body parts grow such that their size ratio remains constant. Miller and Kahn [13] argue in favour of our method thus: 'there is usually no clear justification for saying, e.g. that increase in skull length is dependent upon increase of body length; it is more realistic to consider changes in skull length and body length as due to a set of common factors'. Ricker [16] discusses the value of the method in fishery research. Applications include modelling relationships between weight and length, between weight and fecundity (the number of eggs), and estimating the 'catchability' of fish (the fraction of the stock taken by one unit of fishing effort). Rayner [15] gives an application to the flight speed of birds as related to the windspeed.

We have already noted the scope for application in astronomy. Babu and Feigelson [1] point out that 'differences in regression methods on similar data may be responsible for a portion of the long-standing controversy over the value of Hubble's constant, which quantifies the recession rate of the galaxies'. The earliest appearance of our method in the astronomical literature seems to be that of Stromberg [19].

The method has also been proposed in the context of educational and psychological testing. A very early reference being that of Otis [14] who named it the 'relation line'. If two tests are meant to measure the same aptitude or attainment one may need to match pairs of equivalent scores on the pair of tests for creating a conversion table. The direction of the conversion should obviously not affect which values are paired off, hence the need for a symmetric approach. Greenall [7] proposes the 'equivalence line' for this purpose:

$$\frac{y-\mu_y}{\sigma_y} = \frac{x-\mu_x}{\sigma_x}.$$

This turns out to be yet another name for our least areas line. For standardised scores the line equation reduces to $y = x$. He also proves a very interesting uniqueness result: 'When we seek a relation that will deem a pair of scores mutually equivalent if and only if the proportion of x-scores less than X equals the proportion of y-scores less than Y, we aim at pairing off scores that give rise to equal percentile ranks. In the case of continuous bivariate distributions which satisfy a simple condition $[F(x,y) = F(y/c, cx)]$, only the equivalence relation will provide this relation'. The normal distribution is one case which satisfies this condition. A relevant theoretical result due to Kruskal [12] is that if the two

variables are normally distributed and a line is needed to predict x from y as often as y from x, then the least areas line maximises the probability of correct prediction (i.e. the probability of being within z standard deviations, for any given z-value). This provides another justification for the use of this line.

Hirsch and Gilroy [9] show how it can be useful in hydrology and geomorphology where one may be interested in relationships between e.g. stream slope versus elevation, or stream length versus basin area, etc.. 'In such cases there is no clear direction of causality but there is clearly an inter-relation of variables'. 'A major motivation for the use of the line lies in the equivalence of the cumulative function of y and y_{est}'.

In general terms when should the least areas method be used? Rayner [15] cites the result of Kendall and Stuart [10] that if no error information is available then this method gives the least-bias or maximum likelihood estimate of the functional relation. Rayner goes on to demonstrate that this line also has the property of being independent of the correlation between the errors of the two variables.

Ricker [17] deals with the question of usage by first distinguishing between random measurement error and mutual natural variability (as arises for example in biology). In the former case for each observation there is an associated true point which would arise if the errors in both variables were zero. If one can estimate the variances of the errors by replicating the measurements then measurement error models can be used to estimate the line. One monograph on such models is Cheng and Van Ness [4]. If one cannot estimate the error variances (or their ratio, λ) then Ricker recommends the use of the least areas line as being the best approximation: it gives y and x equal weight and will be exact if $\lambda = \text{var}(y)/\text{var}(x)$, i.e. when the ratio of error variances equals the ratio of data variances. For the case of mutual natural variability 'there is no basis for assigning separate vertical and horizontal components to the deviation', i.e. 'it is impossible to say whether it is y or x that is responsible for the deviations from the line'. In this case Ricker concludes that if the data are binormally distributed then the least areas line be used to describe the central trend, and least squares to estimate one variable from the other. For the mixed case i.e. having both measurement error and natural variability, 'the best that can be done is to treat them in terms of whichever source of variation makes the larger contribution to the total. In biological work this will usually be natural variability'.

Despite appearing in so many other fields, it is remarkable that this technique does not seem to have appeared in the numerical analysis/approximation literature. For example it is not listed in Grosse's Algorithms for Approximation catalogue. The present paper looks at an obvious way of extending the approach to any number of variables by minimising volumes.

3 Least volume fitting

We now intend to fit a linear function of the form $\sum_{j=1}^{p} a_j x_j = c$ to data $\{X_j\}$ in p dimensions, in other words we have data on p variables and we seek a linear relationship between them. Of course this is not uniquely specified because we can divide through by any non-zero constant. Thus we are free to impose a constraint on the coefficients, such

as $c = 1$. Note that we shall not permit any of the coefficients a_j to be zero because that would imply the associated variable x_j is unrelated to the other variables

One obvious way of generalising the least areas procedure to higher dimensions is to minimise the volumes (or hypervolumes). Each data point will have associated with it a 'volume deviation' which is simply the product of its deviations from the fitted plane in each dimension. We must take care to make all these non-negative by taking the absolute values. For the ith data point this volume deviation V_i is proportional to

$$\left| \frac{(\sum_{j=1}^{p} a_j X_{ij} - c)^p}{\prod_{j=1}^{p} a_j} \right|.$$

We now introduce non-negative variables u_i, v_i to deal with the absolute value of the numerator. The positive u_i represent points on one side of the fitted plane, and positive v_i refer to points on the opposite side. Setting $c = 1$ allows us to model the bracketed term thus:

$$u_i - v_i = \sum_j a_j X_{ij} - 1.$$

At least one of each of the pairs $\{u_i, v_i\}$ will be forced to be zero by their presence in the objective function which is being minimised. Consequently the numerator can be represented as $\sum(u_i^p + v_i^p)$. We shall suppose the denominator is positive; if it is not we can always make it so by multiplying one of the x-variables by -1 so that its coefficient, and hence that of the product of coefficients, also changes sign. We can now formulate our problem as the following fractional programme:

$$\begin{aligned} \text{Minimise} \quad & \sum_i (u_i^p + v_i^p) / \prod a_j \\ \text{such that} \quad & u_i - v_i = \sum_j a_j X_{ij} - 1 \\ \text{and} \quad & u_i, v_i \geq 0. \end{aligned}$$

The field of fractional programming is comprehensively covered by Stancu-Minasian [18]. We note that Draper and Yang [6] used a different route to generalising the technique to more than two dimensions. They minimised the pth root of the squared volumes and showed that the estimated coefficients were a convex combination of those from the p OLS estimates.

4 Numerical test

We shall now apply the least volume criterion to try and uncover the coefficients from data that have been generated from a known underlying model with some randomness thrown in. In order to make this a difficult test we shall choose data, which suffers from multicollinearity. This means that there is a near linear dependence within the data, i.e., one of the variables almost lies in the space spanned by the remaining variables, and so we are close to being rank-deficient. The data is taken from Belsley's [3] comprehensive monograph on collinearity. The generating model is

$$y = 1.2 - 0.4x_1 + 0.6x_2 + 0.9x_3 + \epsilon$$

with ϵ normally distributed with zero mean and variance 0.01. The absolute correlations between the variables ranged from 0.35 to 0.61 and so these in themselves would not have alerted the researcher to any difficulty associated with multicollinearity. Two very similar data sets (A,B) are tabulated in Belsley based on this model. For set A ordinary least squares (OLS) gives:

$$y = 1.26 + 0.97x_1 + 9.0x_2 - 38.4x_3.$$

The fit as measured by $\mathbf{R}^2$ is very good at 0.992 but the underlying model is far from being uncovered. In particular, the coefficient of x_2 is 15 times too high and two of the coefficients have the wrong sign! Getting the signs wrong is very serious if one is trying to understand how variables are related to each other. Turning to the least volume approach we find:

$$y = 1.20 - 0.43x_1 + 0.37x_2 + 1.97x_3.$$

We now have all the correct signs and the magnitudes are much closer to the true ones.

Repeating this for data set B:

$$\begin{aligned} \text{OLS}: \quad & y = 1.275 + 0.25x_1 + 4.5x_2 - 17.6x_3 \\ \text{Least volume}: \quad & y = 1.20 - 0.43x_1 + 0.37x_2 + 1.98x_3. \end{aligned}$$

Once again the least volume approach produces a superior model. Moreover it is also worth noting that the two OLS models are very different from each other whereas the least volume models seem to be more stable to small variations in the data. This is noteworthy because of how similar the two data sets were: the y-values were identical for sets A and B, and the x-values never varied by more than one in the third digit. Thus our method seems to be much more stable than OLS. Of course a comprehensive set of Monte Carlo simulations is required to fully explore this aspect.

5 Conclusion

We have presented a fitting method whose criterion combines the deviations in each dimension by multiplying them together. This simple device means that re-scaling of any of the variables e.g. by changing the units of measurement, will give rise to an equivalent model. This property of units-invariance is not shared by the total least squares approach (or orthogonal regression: where the sum of the perpendicular distances to the fitted plane are minimised). By taking the product of the deviations we ensure that all variables are treated on the same basis and this is useful if the purpose is to find an underlying relationship rather than to predict one of the variables.

When we applied the technique to data we were able to recover the underlying relationship much more closely than when least squares was used. Not only were the signs of the coefficient correctly reproduced (which is crucial for understanding directions of change) but also the magnitudes were much closer to the true values than least squares estimates. It appears that the least volume method may be superior when there is multicollinearity in the data. Much more simulation needs to be done to investigate this potentially very valuable feature.

Bibliography

1. G. J. Babu and E. D. Feigelson, Analytical and Monte Carlo comparisons of six different linear least squares fits, *Communications in Statistics: Simulation and Computation*, **21** (2) (1992), 533–549.
2. F. Barker, Y. C. Soh, and R. J. Evans, Properties of the geometric mean functional relationship, *Biometrics* **44**, (1988) 279–281.
3. D. A. Belsley, *Conditioning Diagnostics*, Wiley, New York, 1991.
4. C-L Cheng and J. W. Van Ness, *Statistical Regression with Measurement Error*, Arnold, London, 1999.
5. N. R. Draper and H. Smith, *Applied Regression Analysis* (3rd edition), Wiley, New York, 1998.
6. N. R. Draper and Y. Yang, Generalization of the geometric mean functional relationship, *Computational Statistics and Data Analysis* **23** (1997), 355–372.
7. P. D. Greenall, PD (1949). The concept of equivalent scores in similar tests. *British J. of Psychology: Statistical Section* **2** (1949), 30–40.
8. E. Grosse, (1989). A catalogue of algorithms for approximation, in *Algorithms for Approximation II*, eds. J. C. Mason and M. Cox.
9. R. M. Hirsch and E. J. Gilroy, Methods of fitting a straight line to data: examples in water resources, *Water Resources Bulletin* **20** (5) (1984), 705–711.
10. M. G. Kendall and A. Stuart, *The Advanced Theory of Statistics*, 4th edition, vol.2, 391–409, Griffin, London, 1979.
11. D. K. Kimura, Symmetry and scale dependence in functional relationship regression, *Systematic Biology* **41** (2) (1992), 233–241.
12. W. H. Kruskal, On the uniqueness of the line of organic correlation, *Biometrics* **9** (1953), 47–58.
13. R. L. Miller and J. S. Kahn, *Statistical Analysis in the Geologocal Sciences*, Wiley, NY, 1962.
14. A. S. Otis, The method for finding the correspondence between scores in two tests, *J. of Educational Psychology XIII* (1922), 524–545.
15. J. M. V. Rayner, Linear relations in biomechanics: the statistics of scaling functions, *J. Zool., Lond.(A)* **206** (1985), 415–439.
16. W. E. Ricker, Linear regressions in fishery research, *J. Fisheries Research Board of Canada* **30** (1073), 409–434.
17. W. E. Ricker, Computation and uses of central trend lines, *Canadian J. of Zoology* **62** (1984), 1897–1905.
18. I. M. Stancu-Minasian, *Fractional Programming: Theory, Methods and Applications*, Kluwer Academic, Dordrecht, 1997.
19. G. Stromberg, Accidental and systematic errors in spectroscopic absolute magnitudes for dwarf G0-K2 stars, *Astrophysical J.* **92** (1940), 156–169.
20. G. Teissier, (1948). La relation d'allometrie, *Biometrics* **4** (1) (1948), 14–48.